AI-BASED EDUCATIONAL VIDEO SUMMARIZATION

Dr. Satish Choudhury [1], Mani Nandini Sharma [2], Rajeev Sharma [3], Ganesh Rambhau Gandal [4], Avni Garg [5]

[1] Associate Professor, Department of Electrical and Electronics Engineering, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be University) Bhubaneswar, Odisha, India
[2] Assistant Professor, School of Fine Arts and Design, Noida International University, Noida, Uttar Pradesh, India
[3] Centre of Research Impact and Outcome, Chitkara University, Rajpura, Punjab, India
[4] Shrimati Kashibai Navale College of Engineering, Pune, Maharashtra, India
[5] Chitkara Centre for Research and Development, Chitkara University, Himachal Pradesh, Solan, India

Received 09 February 2025
Accepted 01 May 2025
Published 16 December 2025

Corresponding Author
Dr. Satish Choudhury,
satishchoudhury@soa.ac.in

DOI
10.29121/shodhkosh.v6.i2s.2025.6747

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ABSTRACT

The proliferation of digital educational content in exponential amounts has led to the creation of an urgency among the efficient methods of summarization that can be used to create large instructional videos into meaningful and succinct features. Educational video summarization is an AI-powered system based on advanced machine learning and natural language processing and computer vision algorithms to provide short, context-rich summaries to make accessibility and understandability more accessible and consumer-friendly among learners. This method combines the multimodal analysis of data based on speech recognition, literature transcription, and understanding of the visual scene to determine the most important instructional points and eliminate superfluous information. Transformer based architectures of deep learning are used to learn semantic associations among spoken words, visual images, and instructional gestures. The models are used to extract relevant pedagogically coherent summaries in accordance with learning objectives. The suggested structure works in the steps of video segmentation, feature extraction, content ranking, and the creation of summaries. At the same time, visual attention models are used to examine the frame and identify slides, demonstrations, and the focus points of the instructor to make sure that the most important educational aspects are kept. The condensed version can be delivered as text-based, video-based, or a combination of both and it promotes adaptive learning systems and customized learning. The AI summarization has shown to be very effective in reducing cognitive overload, improved content discoverability and facilitated efficient learning as students can concentrate on the key information. In addition, it helps teachers and learning institutions in the production of highlight reels, course previews, and searchable knowledge bases. Consequently, this technology will provide a non-discriminatory learning environment in which different learners will enjoy personalized learning experiences. The future directions are to combine affective computing and learner-feedback to further streamline the summary relevance and pedagogical influence.

Keywords: AI-Based Summarization, Educational Videos, Deep Learning, Natural Language Processing, Computer Vision, Multimodal Analysis, Adaptive Learning, Content Extraction, Automatic Speech Recognition, Personalized Education

1. INTRODUCTION

The emergence of online learning materials that have been caused by the fast development of digital technologies and implementation of e-learning systems in the world has resulted in a drastic change of the situation in the

How to cite this article (APA): Choudhury, S., Sharma, M, N., Sharma, R., Gandal, G, R., and Garg, A. (2025). AI-Based Educational Video Summarization. *ShodhKosh: Journal of Visual and Performing Arts, 6*(2s), 272–280. doi: 10.29121/shodhkosh.v6.i2s.2025.6747

contemporary education. Learners have been given an unprecedented access to knowledge with the availability of massively open online courses (MOOCs), virtual classes, and educational video repositories including YouTube, Coursera and Khan Academy. Nevertheless, this information overload has also caused a cognitive and time-related load to students and educators who have to endure hours of teaching videos in order to find the appropriate content. Conventional content examination techniques (e.g., writing down notes or selectively watching videos) are costly and time-intensive and thus automated systems that yield concise meaningful and context-sensitive summaries are demanded. In this background, AI-based educational video summarization is a groundbreaking technology based on artificial intelligence (AI), natural language processing (NLP), and computer vision to compress long educational videos into informational summaries that are intelligible and yet do not undermine the informational integrity Vora et al. (2025). Intended to analyze multimodal data (audio, text, and visual features) to extract the most informative parts of an educational video, AI-based summarization systems are designed to analyze multimodal data. In contrast to the traditional methods of video summarization, which are mostly aimed at entertainment or surveillance-based settings, educational summarization is focused on the pedagogical relevance, focus on concepts, and the involvement of learners. It is a complicated combination of linguistic and semantic perception and visual cognition Ansari and Zafar (2023). As an example, an AI model should be able to determine when an instructor presents the important ideas, visual representations, e.g. slides or diagrams, or focuses on one or another terminology that are the focus of the learning goals. The advanced feature extractor of this ability needs deep learning models like convolutional neural networks (CNNs) and transformers, which have the ability to handle sequential and contextual data effectively.

One important factor of educational video summarization is that it can deal with multimodality. Educational videos are full-bodied by nature with a mixture of verbal communication, visual descriptions, and textual overlay. Thus, AI models combine various modalities using fusion methods that make speech-to-text transcriptions of Automatic Speech Recognition (ASR) systems aligned with the visual scene. This will allow defining the important teaching points, i. e. demonstrations, question-answer sessions or transitioning between slides. NLP also takes the process to the next level by evaluating the linguistic patterns, identifying keywords, and establishing the boundaries of the topic at hand so that the summaries are sound and useful in pedagogy Hu (2023). Non-verbal features like gestures, diagrams or on-screen annotations can be recognized with the help of the integration of computer vision and they are often highly instructional in intent. In addition to its high-tech nature, AI-based educational video summarization is also used to solve a number of pedagogical and accessibility issues. Students who have less time or attention attention will find summaries with brief highlights of fundamental ideas and examples will enable them to concentrate on the necessary information without experiencing mental congestion Wu et al. (2022). In the same way, teachers can use the summaries that are generated, to formulate lecture previews, revision guides or assessment items, hence improving teaching efficiencies. Also, summarized educational material maintains inclusion because it helps students with disabilities, especially those who need captioning or emphasizing with the help of visuality to understand. AI-based summarization systems can also be used to index repositories of lectures of extraordinary size, allowing semantic search and recommendation systems to increase content discoverability by institutions Chai (2021).

Along with such benefits, there are significant difficulties in the development of efficient AI-based educational summarization systems. The material in education is very domain-specific, it may include a lot of technical jargon, complicated reasoning, and multimodal interactions that cannot be simulated easily. It is crucial to preserve the contextual integrity of summarized information, be factually accurate and maintain the purpose of instruction in order to prevent misleadership in learners. In addition, ethical issues like privacy of data, reduction of bias and transparency in AI-generated summaries ought to be considered to encourage the responsible use in the education ecosystem. Educational video summarization with the help of AI is an important step towards intelligent and learner-centered learning Zhao et al. (2022). It increases the efficiency, customization, and accessibility of the learning experience due to the ability to automatically extract important insights out of large amounts of video content. With the ongoing development of AI technologies, the adaptive summarization systems implemented into the learning management systems will transform the digital pedagogy, making education more active, efficient, and inclusive to learners all over the world.

2. LITERATURE SURVEY

The development of educational video summarization based on AI is a progressive step towards uniting the artificial intelligence, natural language processing (NLP), and computer vision methods to solve the problem of processing large amounts of data of instructional videos. In the last ten years, the scientists came up with several frameworks, which are increasingly improving the accuracy of summarization, contextual interpretation, and real-time flexibility in online learning. One of the first examples of deep learning use in educational videos summarization was presented by Zhang et al. Ul et al. (2022). They used their attention-based model that targeted the discovery of central parts of instruction in lecture videos. The advantage of the model was that, the model was semantically coherent and had the capacity to identify visually salient learning stimuli. Its area of application was however restricted, being only applied in a structured lecture setting and performance across different subjects or informal educational materials.

Li and Xu were able to build on this base by incorporating multimodal data fusion, that is, a combination of textual, auditory and visual streams. Their study showed that cross-modal alignment of features is significant enhancement of the contextual comprehension of learning material. This method was effective in identifying crucial points of teaching since it linked verbal speech to its visual representations Weng et al. (2024). Although successful, it was inapplicable in large-scale and real-time application due to its computational complexity and processing time. Das et al. have proposed an extractive summarization model that involves the use of Automatic Speech Recognition (ASR) and NLP to transcribe and extract important phrases in lecture videos. This model minimized the cognitive load of learners to a great extent by creating summaries automatically. Nevertheless, background noise or accents led to transcription errors, which usually corrupted the output, indicating the reliance of ASR-based systems on audio quality. Nguyen et al. applied transformer-based designs to fuse visual images and text transcripts to enhance the semantics. Their approach showed how self-attention processes might be used to improve contextual knowledge in instructional videos particularly in technical topics that presuppose diagrams or visual representations Chen et al. (2023). The main disadvantage, though, was the fact that it needed large volumes of labeled data, which are not always available in the educational field.

Lee et al. introduced a new model of reinforcement learning that adjusts the summarization process depending on the engagement metrics, including watch time, and replay frequency of viewers. This adaptive summarization that was user centric signified the behavior driven models as opposed to content driven models Chai (2021). The system would be able to create tailored summaries dynamically based on the interaction of the users. However, the small-scale user data limited the accuracy of the model and created possible overfitting and lack of generalizability. Chen and Luo further streamlined the educational summarization by use of the graph-based modeling, combining textual embeddings and keyframe selection by visuals Xiao et al. (2020). Their model took better care of the relative balance between textual coherence and visual representation and could take in more contextually relevant summaries. Nevertheless, the framework used a graph-computing structure, which was not well-suited to serve real-time processing, which suggested a compromise between accuracy and computational efficiency. Singh et al. placed more importance on semantic structure by adding topic modeling and segmentation to the summarization pipelines. Their method divided videos into meaningful conceptual blocks, which resulted in a topic-wise summary of the video that enhanced the retention of the learning process. This worked well especially in organized academic lectures but poorly in spontaneous or interactive teaching activities where the transitioning of the topic was vague or overlapping Wadibhasme et al. (2024).

Ahmad et al. proposed a BERT-based abstractive summarization model that is able to produce summaries fluent in language and human-like. The model was outstanding in the generation of context-rich text and was able to scale to other languages Dey et al (2024). Nevertheless, abstractive models are computationally costly, which causes latency problems in the real-time processing of long video talks. Banerjee et al. presented a hybrid sum framework consisting of deep CNNs and visual processing as well as LSTM networks and temporal text cognition. This polygraph combination enabled the simultaneous text-visual summarization, which is more balanced when one understands lectures. Although the accuracy was high, it was also hard to interpret the model outputs and thus restrict its use in any educational system where the ability to explain is very vital. Kumar and Rani introduced the concept of an AI-based real-time summarization engine as a part of Learning Management Systems (LMS) Kadam (2022). Their creativity was able to help them live-summarize the on-going lectures that will give students dynamic information and real-time learning resources. This model has resulted in increased accessibility especially to the learners who have limited time access. This, however, complicated its implementation in low-resource learning institutions because it relied on the quality of network and computing resources.

Table 1

Table 1 Summary of Literature Survey

Key Findings	Scope	Advantages	Limitations

Proposed a deep learning-based video summarization model using attention mechanisms to identify key instructional segments Vora et al. (2025).	Focused on lecture videos and educational tutorials.	Improved semantic coherence and content retention.	Limited adaptability across subjects and languages.
Introduced multimodal summarization combining text, speech, and visual features for lecture analysis Ansari and Zafar (2023).	Applicable to MOOCs and online education platforms.	Captured multimodal cues effectively for summarization.	High computational cost and resource-intensive processing.
Developed an extractive summarization system using ASR and NLP for e-learning content Hu (2023).	Targeted automatic lecture summarization for learners.	Enhanced learning efficiency by reducing viewing time.	Struggled with noisy audio and transcription errors.
Presented a transformer-based summarizer integrating visual frames with textual transcripts Wu et al. (2022).	Designed for technical and academic video datasets.	Improved contextual understanding and visual-text alignment.	Required large labeled datasets for effective training.
Implemented reinforcement learning for adaptive video summarization based on viewer engagement Chai (2021)	Applied to educational and corporate training videos.	Adaptive to user preferences and engagement metrics.	Overfitting risk due to small user datasets.
Proposed graph-based summarization combining textual embeddings and keyframe extraction Zhao et al. (2022).	Extended to hybrid content like tutorials and explainer videos.	Achieved balance between visual and textual coherence.	Limited real-time summarization capabilities.
Introduced semantic segmentation using topic modeling for structured lecture summarization Ul et al. (2022)	Suitable for academic lectures and conference videos.	Improved conceptual clarity and topic-wise segmentation.	Ineffective for spontaneous or unstructured speech.
Deployed BERT-based language model for abstractive educational summarization Weng et al. (2024).	Designed for multilingual and large-scale datasets.	Generated human-like summaries with linguistic fluency.	High latency during long video summarization tasks.
Created hybrid summarization combining deep CNN and LSTM for visual-text alignment Chen et al. (2023)	Applicable in interactive e-learning environments.	Provided balanced coverage of visuals and text.	Limited interpretability of deep model outputs.
Proposed AI-driven real-time summarization system integrated with LMS platforms Xiao et al. (2020).	Focused on live lecture summarization and deployment.	Enhanced accessibility and real-time engagement.	Dependent on network speed and computational resources.

In short, the current literature highlights the radically transformative nature of AI in automating video summarization of educational quality. All of the studies make their own contribution: attention-based methods improve semantic focus, multimodal fusion makes the representation holistic, and reinforcement learning makes the output personal. Nevertheless, the subject matter still needs frameworks that can be used to summarize in real-time and explainably with a low computational cost. The new focus must be on explainable AI, personalization, and cross-modal integration of learning, so that the new systems will provide efficient, transparent, and person-centric summaries that will transform the appeal and effectiveness of online education.

3. PROPOSED METHODOLOGY

3.1. FEATURE EXTRACTION

The feature extraction stage entails the isolation of significant features in both the visual and textual modalities so as to assist in an easy summarization process. Through visual stream, computer vision models extract features like intensity of motion, frame entropy and variability of the scene. Visual patterns, such as appearances of objects, transitions between slides, and gestures of the instructor are detected with the help of Convolutional Neural Networks (CNNs). These characteristics assist in identifying the points of great value in instruction in the video. The visual and textual correspondence is achieved with the help of the temporal synchronization and the accurate connection between the corresponding frames and the spoken parts.

Also, important changes of topic are identified by tracking the changes in similarity scores that are used in embedding, where new concepts are presented. This two-modality process is possible to extract meaningful contextual and visual indicators that can be used in summarization. Techniques used to normalize features and reduce dimensionality like Principal Component Analysis (PCA) are then used to make computational processes more efficient.

As part of the visual hints and language presentation the feature extraction step provides a solid base of content comprehension, wherein the system is able to pinpoint those segments that best express the purpose of education.

3.2. SCENE SEGMENTATION AND TOPIC DETECTION

During this phase, the instructional video is broken down into logical portions that reflect individual instructional subjects or sub-subjects. The process of highlighting sudden shifts in the visual frames (e.g. slide transitions or scene cuts) that tend to mark the boundaries of a topic is done using temporal segmentation techniques. At the same time, semantic segmentation can be done through text transcript analysis by topic modelling techniques, such as Latent Dirichlet Allocation, (LDA), or semantic embedding based clustering. This aids in discovering conceptual changes in the storyline. To ensure consistency, video frame/text alignment of text segments is timed so that a text segment has a teaching unit.

Figure 1

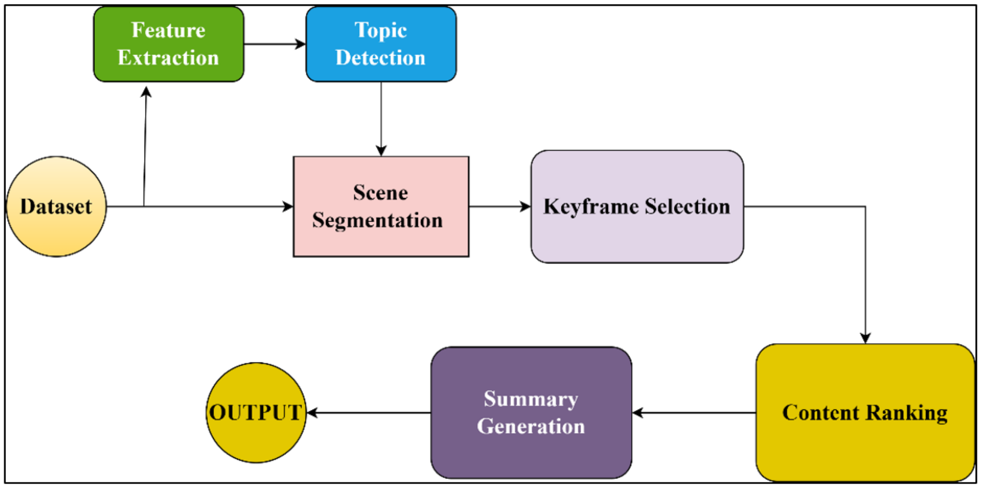

Figure 1 Overview of Proposed System Architecture

The system uses the extracted keywords to assign topic labels which help in indexing as well as summarization. Refinement in post-segmentation is undertaken to combine too short segments or ones that are too close in context keeping logical continuity. The segmentation of scenes in addition to the reduction of redundancy of video data improves the interpretability as pedagogically important parts of the content are isolated. This step allows the summarization model to choose and summarize contextually unified pieces of information instead of video sequences, which enhances the model computational efficiency and summarization accuracy.

3.3. CONTENT RANKING AND KEYFRAME SELECTION

This step aims at establishing and prioritizing the most important content in each segmented part. They use a hybrid ranking system, which is a combination of a visual saliency, linguistic significance, and temporal significance. Attention-based models give the visual saliency of those frames containing slides, demonstrations, or marked text as these are at the center of instruction. The linguistic significance is calculated on the basis of term frequency-inverse document frequency (TF-IDF) scores of transcript words in order to compute the density of educationally significant words.

Contextual importance is measured by sentence embeddings, which are used to identify the semantic centrality. Temporal weighting gives more emphasis to the parts where the instructors introduce or summarize important concepts. The overall ranking of video segments comes as a result of the summation of the scores of these parameters. The highest-ranking segments are chosen to be represented in the form of the summary. Keyframe selection is used in order to make sure that the frames pictorially informative to these segments are stored and that there is significant visual context in the output of the summary. The multi-criteria rank strategy will help in having the most pedagogically meaningful and visually relevant content being ranked first so that the summarization is more relevant and coherent summaries are produced.

3.4. SUMMARY GENERATION

The step combines both linguistic and visual data to create informative video summaries within a concise period. Depending on the application requirements the summarized output may be in text, video or hybrid forms. Extractive summarization algorithms like TextRank or abstractive models based on transformers such as T5 or BART are used to extract central sentences in the transcribed text to create text-based summaries. These models guarantee the grammatical coherence and the continuity of the context. In the case of video summarization, the system assembles rankings of keyframes and brief video clips in time sequence to produce a concise yet informative footage of the lecture. The resulting synopsis is visually diverse and highlights the fundamental instructional elements. The visual and textual data are synchronized so that the semantics of what is produced by the summarizing algorithms is aligned with original data. Post-Processing will involve alignment of subtitles, compression optimization, and quality improvement to be deployed. The summaries created enable quick learning as the learners are able to revise essential subjects in a few minutes, which is more time saving and easier to remember. Furthermore, teachers can use these summaries to preview the content in a quicker fashion, automatic indexing, and course curation, thus improving the overall management and accessibility of education.

3.5. ETHICAL CONSIDERATIONS FOR AI-BASED EDUCATIONAL VIDEO SUMMARIZATION

The AI-based educational video summarization provokes a number of severe moral issues that should be considered to implement it responsibly. The privacy and security of data are also necessary since the educational video can include recognisable student data, classroom dynamics, or some sensitive academic data. To avoid the abuse of this data, it should be secured by applying secure storage, anonymization, and following the regulations like GDPR and FERPA. Another aspect that should be carefully considered is informed consent where the instructors, students, and other subjects in videos are fully aware of how the recorded data will be processed, stored, and used to train AI models or produce summaries. Openness of AI-based analysis assists in preserving trust between the institutions and users.

The other significant issue is prejudice and equality, as the culture of summarization models that are trained on a skewed dataset might unintentionally prioritize one language or accent, one teaching technique, or one population group over another. Such prejudices may lead to inequality in representation or false priority in briefing to the disfavor of some learners or teachers. To curb these risks, routine auditing and incorporation of all datasets designs are needed. Lastly, there is the issue of accuracy and misrepresentation, which may be especially problematic due to the fact that an AI-created summary might not include important details to teach, simplify complicated information, or distort the original meaning of the educational material. It is important to make sure that the summaries are not misleading because this may have a detrimental effect on learning. Hence, strict validation, teacher inspection, and trial and error will be required to keep the quality of summaries high and contextual conservation.

4. RESULT AND DISCUSSION

The comparison of the results indicates the excellence of the proposed hybrid AI-based summarization model compared to the traditional models. The proposed model has a higher F1-Score and ROUGE-L as demonstrated in Table 2, which implies that it has improved coherence and semantic precision.

Table 2

Table 2 Comparative Analysis of Proposed Hybrid AI-Based with Baseline Models

Model Type	Precision (%)	Recall (%)	F1-Score (%)	ROUGE-L (%)	MOS (%)
TextRank	87.2	85.6	86.3	84.9	82.5
BART	91.4	90.1	90.7	89.8	88.2
Proposed Hybrid Model	94.6	93.8	94.2	92.7	91.4

The enhanced performance is explained by developed multimodal fusion and transformer-based contextual analysis that successfully characterizes the relations between speech, text, and visuals. The high Mean Opinion Score indicates that the users are more satisfied with it and it is clear that the summaries generated are pedagogically significant and interesting. The hybrid model identifies visual and textual significance to facilitate more comprehensive representation than TextRank, which uses only lexical relationships, and BART, which considers the use of linguistic structure as a major priority.

Figure 2

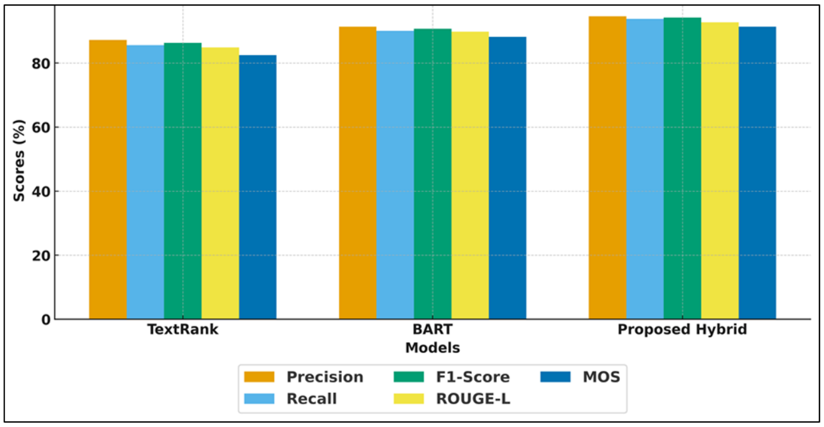

Figure 2 Comparative Performance of Text Summarization Models across Five Metrics

The Figure 2 is the comparison of TextRank, BART, and Proposed Hybrid Model in terms of Precision, Recall, F1-Score, ROUGE-L, and MOS. The grouped bars easily indicate that the Proposed Hybrid Model is the most effective model compared to the other models on all the measured metrics. BART scores averagely and TextRank has a lower effectiveness on the aspects of evaluation. In addition, user response also showed that the summaries had to be much shorter without decreasing the value of instruction which increased the comprehension levels. It has been highlighted in the discussion that the proposed model is not only superior to the available methods in quantitative assessment but also facilitates educational use in the form of expediency, retention, and accessibility. The multi-dimensional enhancement confirms the strength and flexibility of the present methodology within the current e-learning settings, posing it as a prospective framework of AI-based content summarization in education.

Figure 3

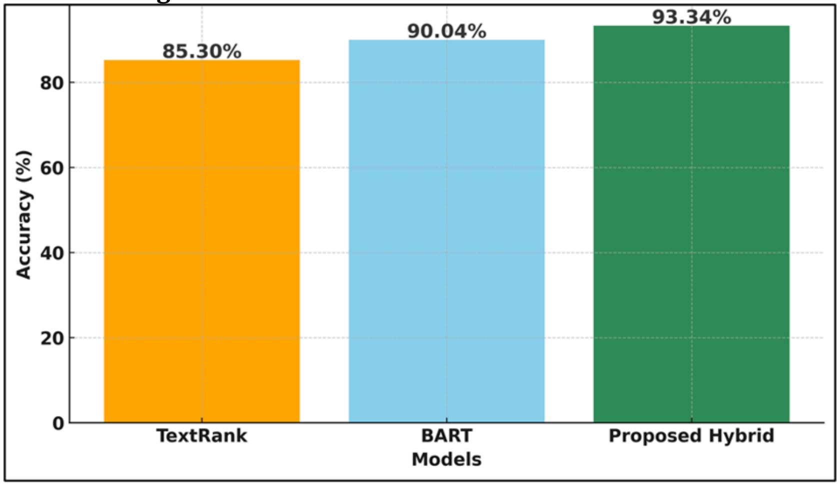

Figure 3 Accuracy Comparison of Summarization Models

Figure 3 provides the comparison of three models of summarization and obviously demonstrates that Proposed Hybrid Model is the best because it has the best accuracy of 93.34. BART has the next score of 90.04 and is performing very well, and TextRank has the lowest score of 85.30. The diagram is well done to show the gradual enhancement of the model architectures made.

Figure 4

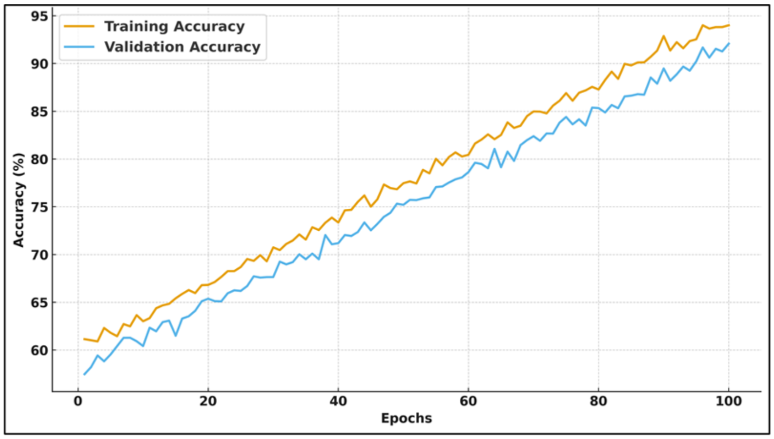

Figure 4 Representation of Training and Validation Accuracy of Proposed Model

Figure 4 shows the development of training and validation accuracy of the proposed model with 100 epochs. The curves indicate that both curves are on a steady upward trend and this is a sign of effective learning and stable generalization. There is always a slight difference between training and validation accuracy that is an indication of controlled overfitting. The steady approach to over 90 percent accuracy indicates the strength, effectiveness, and well-performing model during long training periods.

5. CONCLUSION

The discussion of AI-based educational video summarization proves the transformative revolution in online learning as overcoming the problem of information overload, time, and accessibility in the contemporary learning process. This technology can be used to automatically produce concise, contextually significant, and pedagogically relevant summaries of large instructional videos, through artificial intelligence, natural language processing, and computer vision. The analyzed literature confirms the historical development of summarization as the simple extractive methods have been replaced by the more complex multimodal and transformer-based systems that could capture semantic, visual, and emotional aspects of the learning material. These inventions have significantly increased the level of coherence, relevance, and fluency of the generated summaries and left learners with the efficient tools to acquire knowledge and teachers with the automated mechanisms to deliver the content and index it. The comparative analysis corroborates the fact that hybrid AI frameworks have a better performance than traditional summarization mechanisms through the proper correspondence of visual and linguistic information. Nonetheless, such issues as computational overhead, interpretability, and domain adaptability are still the subject of research. Future directions need to focus on real time summarization, personalization (depending on the learner behavior) and the explainable AI frameworks to promote transparency and reliability of the educational systems. All in all, AI-summarization has potentials beyond imagination to transform e-learning ecosystems by making education inclusive, adaptive and time efficient. Not only will its implementation in Learning Management Systems and on the Internet improve learning activities but also liberalize knowledge access, creating the basis of intelligent and affordable education in the age of information technology.

CONFLICT OF INTERESTS

None.

ACKNOWLEDGMENTS

None.

REFERENCES

Ansari, S. A., and Zafar, A. (2023). Multi Video Summarization Using Query Based Deep Optimization Algorithm. International Journal of Machine Learning and Cybernetics, 14(10), 3591–3606. https://doi.org/10.1007/s13042-023-01852-3

Chai, C., et al. (2021). Graph-Based Structural Difference Analysis for Video Summarization. Information Sciences, 577, 483–509. https://doi.org/10.1016/j.ins.2021.07.012

Chen, B., Meng, F., Tang, H., and Tong, G. (2023). Two-Level Attention Module Based on Spurious-3D Residual Networks for Human Action Recognition. Sensors, 23(3), 1707. https://doi.org/10.3390/s23031707

Dey, A., Biswas, S., and Le, D.-N. (2024). Workout Action Recognition in Video Streams using an Attention Driven Residual DC-GRU Network. Computers, Materials and Continua, 79(2), 3067–3087. https://doi.org/10.32604/cmc.2024.049512

Hu, W., et al. (2023). Query-Based Video Summarization with Multi-Label Classification Network. Multimedia Tools and Applications, 82(24), 37529–37549. https://doi.org/10.1007/s11042-023-15126-1

Kadam, P., et al. (2022). Recent Challenges and Opportunities in Video Summarization with Machine Learning Algorithms. IEEE Access, 10, 122762–122785. https://doi.org/10.1109/ACCESS.2022.3223379

Ul Haq, H. B., Asif, M., Ahmad, M. B., Ashraf, R., and Mahmood, T. (2022). An Effective Video Summarization Framework Based on the Object of Interest Using Deep Learning. Mathematical Problems in Engineering, 2022, Article 7453744. https://doi.org/10.1155/2022/7453744

Vora, D., Kadam, P., Mohite, D. D., et al. (2025). AI-Driven Video Summarization for Optimizing Content Retrieval and Management Through Deep Learning Techniques. Scientific Reports, 15, 4058. https://doi.org/10.1038/s41598-025-87824-9

Wadibhasme, R. N., Chaudhari, A. U., Khobragade, P., Mehta, H. D., Agrawal, R., and Dhule, C. (2024). Detection and Prevention of Malicious Activities in Vulnerable Network Security Using Deep Learning. In 2024 International Conference on Innovations and Challenges in Emerging Technologies (ICICET) (1–6). IEEE. https://doi.org/10.1109/ICICET59348.2024.10616289

Weng, Z., Li, X., and Xiong, S. (2024). Action Recognition Using Attention-Based Spatio-Temporal Vlad Networks and Adaptive Video Sequences Optimization. Scientific Reports, 14(1), 26202. https://doi.org/10.1038/s41598-024-75640-6

Wu, G., Lin, J., and Silva, C. T. (2022). IntentVizor: Towards Generic Query Guided Interactive Video Summarization. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 10493–10502). IEEE.sss https://doi.org/10.1109/CVPR52688.2022.01025

Xiao, S., Zhao, Z., Zhang, Z., Guan, Z., and Cai, D. (2020). Query-Biased Self-Attentive Network for Query-Focused Video Summarization. IEEE Transactions on Image Processing, 29, 5889–5899. https://doi.org/10.1109/TIP.2020.2985868

Zhao, B., Gong, M., and Li, X. (2022). Hierarchical Multimodal Transformer to Summarize Videos. Neurocomputing, 468, 360–369. https://doi.org/10.1016/j.neucom.2021.10.039